

Asociación de Historia Contemporánea
Actas del XIV Congreso

DEL SIGLO XIX AL XXI. TENDENCIAS Y DEBATES
(Alicante, 20-22 de septiembre de 2018)

Mónica Moreno Seco (coord.)
Rafael Fernández Sirvent y Rosa Ana Gutiérrez Lloret (eds.)



**BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES**
www.cervantesvirtual.com

Biblioteca Virtual Miguel de Cervantes
Alicante, 2019

Asociación de Historia Contemporánea. Congreso (14.º. 2018. Alicante)

Del siglo XIX al XXI. Tendencias y debates: XIV Congreso de la Asociación de Historia Contemporánea, Universidad de Alicante 20-22 de septiembre de 2018 / Mónica Moreno Seco (coord.) & Rafael Fernández Sirvent y Rosa Ana Gutiérrez Lloret (eds.)

Alicante, Biblioteca Virtual Miguel de Cervantes. 2019. 2019 pp.

ISBN: 978-84-17422-62-2

Alicante, Biblioteca Virtual Miguel de Cervantes, 2019.

Este libro está sujeto a una licencia de “Atribución-NoComercial 4.0 Internacional (CC BY-NC 4.0)” de Creative Commons.



© 2019, Asociación de Historia Contemporánea. Congreso

Algunos derechos reservados

ISBN: 978-84-17422-62-2

Portada: *At School*, Jean-Marc Côté, h. 1900.

LA MEMORIA DIGITAL DE ESPAÑA. EL ARCHIVO WEB COMO NUEVA FUENTE PARA LA HISTORIA DEL PRESENTE

Francesco D'Amato
(Universidad de Valencia)

Introducción

El fenómeno del web *archiving* es relativamente noto. A menudo se mencionan herramientas como el más famoso depósito de recursos digitales del mundo, Internet Archive, pero también otras como PANDORA en Australia o PadiCat en Cataluña. Se trata de la memoria de la red que, paulatinamente, se está fragmentando en las memorias nacionales de la red. En este contexto se inserta el Archivo de la Web Española (AWE), de la que existen ya algunas descripciones, especialmente de la mano de sus conservadores. Sin embargo, no se ha analizado cómo los investigadores y, en particular, los historiadores, se pueden aproximar a esta herramienta, aún en desarrollo. Aquí, pues, queremos resaltar algunos de los problemas culturales, legales, económicos y técnicos vinculados a la creación, el mantenimiento y el acceso de los usuarios a al AWE, un servicio ofrecido hoy en día por la Biblioteca Nacional de España (BNE). Este servicio mira a limitar una de las propiedades negativas de la información digital en la Red: la volatilidad, es decir, la continua desaparición de las fuentes en Internet tras su consulta (Eiroa, 2018: 91).

Ya a principios del siglo XXI, se percibía internet como el archivo multifacético de la memoria global. Sorprendía la posibilidad de encontrar información actualizada y de almacenar contenidos heterogéneos accesibles en cualquier momento y lugar. Las páginas web podían ser instantáneas y dinámicas, pero se revelaron inestables y efímeras. En 1999, era posible encontrar poco más de la mitad de los sitios web existentes el año anterior (Lyman, 2002). Como afirmaban los historiadores a principios de los 2000, el Web carecía de un carácter esencial para que se le pudiera considerar un archivo: la garantía de encontrar, tras un tiempo razonable, lo que allí se había almacenado (Vitali, 2004: 208). Desde el principio, la exigencia de conservar estos contenidos chocó con una contradicción técnica y cultural: ¿qué sentido tiene intentar capturar unos documentos que, por su naturaleza, mutan continuamente y están interconectados entre ellos?

En este texto, en primer lugar, haremos un breve repaso a los proyectos y a la literatura en materia de archivado web, mencionando también algunas nuevas reflexiones de la historia del presente y de las humanidades digitales. En segundo lugar, hablaremos de los recursos tecnológicos necesarios para archivar los sitios web con todos sus recursos. En tercer lugar, a través de un ejemplo de posible nueva colección del AWE, sobre historia contemporánea e historiadores, analizaremos las tareas necesarias para empezar y mantener una colección. Finalmente, intentaremos llegar a unas conclusiones que son inevitablemente el fruto de la mezcla entre mi formación como historiador de la España contemporánea, la pasión para el mundo de las bibliotecas y del patrimonio documental y de unas fructuosas prácticas como conservador web en la BNE realizadas entre mayo y agosto de 2017.

El objetivo es reflexionar sobre la necesidad de acercar a los usuarios a esta herramienta, no sólo para usarla como fuente para la investigación, sino para que comprendan la creación y la gestión de las colecciones y puedan contribuir, de formas directas e indirectas, a su desarrollo.

Una breve revisión de los casos y de los estudios sobre el web *archiving*

«Archivar la web consiste en seleccionar, capturar, almacenar, preservar y habilitar el acceso futuro a páginas y recursos de Internet», como ha sintetizado Mar Pérez Morillo (2014: 125), Jefe de Área de Gestión del Depósito de las Publicaciones en Línea de la BNE. El archivado web no se limita a fotografiar una página web, sino que trata de conseguir un reflejo de la posibilidad de navegación dentro del sitio web y de interconexión con otros sitios a través de sus hiperenlaces. Esta tarea ha crecido paralelamente al desarrollo de Internet y su importancia se ha difundido paulatinamente entre los investigadores en todos los continentes.

Internet Archive, una organización no gubernamental estadounidense, y la National Library of Australia se han impuesto como las entidades de conservación pioneras y líderes en el mundo, ya que sus actividades empezaron en 1996 y sus archivos web son de los más conocidos y utilizados. Esto porque sus archivos son casi del todo públicos y, como en los casos de los archivos web de la Library of Congress (activo desde 2000) y de la Bibliothèque Nationale de France (desde 2002), hacen recolecciones masivas y selectivas de eventos relacionados con el mundo entero y con sus propios países. Desde entonces, diversas bibliotecas patrimoniales han empezado a recolectar, de forma masiva o selectiva, los contenidos de la Red⁸⁰⁵.

En España, Cataluña y País Vasco empezaron a recolectar su patrimonio digital comunitario respectivamente en 2005 y en 2007. El PADICAT, acrónimo de la iniciativa de la Biblioteca Catalana para la conservación del patrimonio digital de Cataluña, se encuentra en un estado muy avanzado y está desarrollando sus propias herramientas. ONDARENET, la iniciativa vasca para la creación del Euskal Ondare Digitala, intenta recolectar, como en el caso catalán, el patrimonio digital vasco, dividiéndolo según las áreas del conocimiento humano⁸⁰⁶.

También gracias a estos modelos dentro y fuera de España, la BNE se ha sumado al archivado web desde 2009: ha entrado en diversos grupos de colaboración internacionales y ha creado una red de conservación del patrimonio digital nacional que involucra las más importantes bibliotecas regionales del país, con la exclusión de Cataluña y País Vasco (Muguerza y Pérez, 2012). Los archivos y las bibliotecas han asumido esta tarea como «una extensión de su misión de preservación del patrimonio bibliográfico» (Pérez, 2015). Además, si la BNE tuviera la posibilidad de recoger unos sitios web de otras entidades que representan parte del patrimonio documental de España, debería acogerlos, analizarlos e integrarlos en su propio archivo de la web, tal y como hizo la Stanford University con las páginas web de los años 90 de SLAC (AlSum, 2016).

Las «Directrices para la preservación del patrimonio digital» de la UNESCO de 2003, diseñadas por la National Library of Australia, pionera del archivado web, demostraron el interés a nivel

⁸⁰⁵ El listado de los archivos web en el mundo se puede consultar en la Wikipedia en inglés: aunque no esté actualizada destaca que se trata de iniciativas vinculadas especialmente con el norte del mundo (con excepciones lamentables como en el caso de Italia, Irlanda y el este europeo) Disponible en: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

⁸⁰⁶ L'Arxiu Web de Catalunya permite búsqueda por texto completo y URL y está organizado por colecciones, directorios temáticos y monografías. Disponible en: <http://www.padicat.cat>. En el caso vasco, se permiten incluso búsquedas avanzadas y la elección de los formatos. Disponible en: <http://www.ondarenet.kultura.ejgv.euskadi.eus:8085/ondarenet/>.

global en la conservación del patrimonio bibliográfico en línea. A ella siguieron las diversas legislaciones nacionales y la creación de un espacio de colaboración llamado International Internet Preservation Consortium (IIPC). Este Consorcio, en 2003, reunía once bibliotecas nacionales y una fundación sin ánimo de lucro, Internet Archive, el archivo de Internet libre y gratuito más extenso y famoso. El IIPC es ahora considerado el organismo encargado, por un lado, de permitir el dialogo entre los representantes de casi 50 países para mejorar y desarrollar el archivado web y, por otro, de empujar la normalización de las normas y de las tecnologías vinculadas a ello (Goethals *et al.*, 2015)⁸⁰⁷.

En España, la Ley de Depósito legal (23/2011, 29 de julio) reconoció que también las publicaciones electrónicas son parte del patrimonio documental y, por esta razón, están sujetas al depósito legal para su preservación y difusión. Esto significa que se amplía a estos nuevos recursos la obligación de conceder un ejemplar, en cualquier soporte, de todas las publicaciones españolas. El Real Decreto sobre el Depósito legal de las publicaciones en línea (635/2015, de 10 de julio)⁸⁰⁸ confirmó que también las informaciones contenidas en los soportes electrónicos no tangible, pero identificables, como los sitios web, se considerarían parte de las publicaciones con obligo de depósito en las instituciones designadas para archivar y difundir estos contenidos con fines culturales o de investigación. Todas las páginas web producidas por personas o entidades españolas (o residentes en España), en cualquiera de los idiomas oficiales, se considerarían objeto de depósito legal, independientemente del servidor y del dominio, es decir, no sólo los «.es». La legislación excluye la correspondencia privada y todos los contenidos destinados a una red privada o con acceso restringido a un grupo determinado de personas (González, 2015).

Si en el caso de los patrimonios documentales en soporte materiales son los editores que se encargan de depositar los ejemplares de las obras, en el caso del depósito legal electrónico la iniciativa no está en los editores. Por esta razón, es la BNE que se encarga de seleccionar y establecer las prioridades de rastreo, mientras que los editores están obligados a asegurar el acceso al robot.

Pese a las numerosas iniciativas, el contenido de Internet es inabarcable: se puede sólo mitigar esta tendencia a la merma de páginas web y las capturas nunca consiguen la total exhaustividad. Tanto desde el punto de vista del número de sitios web archivados, como por la escasa profundidad o número de documentos recolectados para cada sitio web, los esfuerzos de los centros de conservación no han podido evitar la pérdida de la gran mayoría de los contenidos en la Red (Barone, Zeitlyn y Maer-Schönberger, 2015).

Desde finales del siglo XX, se empezó a hablar de lo que se ha definido digital *dark ages* (Wall, 2015). Los historiadores, en particular, entendieron que se estaba asistiendo a un cambio determinante en la producción de los documentos de todo tipo. El pasaje, no sólo de los documentos administrativos, del papel a los diferentes formatos digitales, empujó los investigadores a interrogarse sobre la posibilidad de investigar en un futuro en el que esos documentos digitales habrían desaparecido o serían inutilizables por la obsolescencia tecnológica. Aún hoy en día el peligro no parece conjurado, especialmente en los países que no han activado medidas para archivar su patrimonio digital (Weiner, 2016).

⁸⁰⁷ Véase las páginas web oficiales de estas entidades: <http://netpreserve.org/> y de <https://archive.org/web/>.

⁸⁰⁸ Esta medida legislativa es fundamental para consolidar la labor del archivado web en España. Disponible en: http://www.boe.es/diario_boe/txt.php?id=BOE-A-2015-8338.

Las inquietudes para evitar la extensión de esta edad oscura causada por la obsolescencia digital o la eliminación prematura de ciertos contenidos se han difundido entre bibliotecarios, archiveros, historiadores, sociólogos, politólogos, etc., que se han interrogado sobre la utilidad de archivar la web y la importancia que puede tener en el estudio del presente y en la interpretación del pasado (Webster, 2016). Ejemplos de esta reflexión son algunas recientes obras colectivas especialmente del mundo anglosajón y europeos (Brügger y Schroeder, 2017).

La literatura sobre el depósito legal electrónico y las herramientas y las funciones de los archivos web es relativamente reciente. Por lo general, no tiene más de 15 años y los artículos publicados en las revistas especializadas y las entradas en los blogs sobre bibliotecas se han multiplicado en los últimos 5 años. Para empezar, hay que recordar el Informe Técnico ISO/TR 14873:2013 que funciona como un estado de la cuestión en materia de archivos web. El informe, analizado por Mar Pérez (2014), quería señalar algunas directrices básicas, métodos de recolección, datos estadísticos, indicadores para verificar la calidad de los rastreos y aportar reglas y sugerencia a todos los centros dedicados a esta tarea.

Por otro lado, el archivado web se ha convertido en un tema privilegiado para los especialistas en bibliotecas y archivos, como demuestra la publicación de algunas tesis doctorales sobre esta materia (Banos, 2015). Estas evidencian que, no sólo la documentación producida por los organismos públicos o los individuos han migrado hacia la versión digital en la web, sino que, gracias a esta migración, la documentación vive en otros contextos y ha adquirido características diferentes (Chebbi, 2013). Los mismos conservadores web -así se llaman los que se preocupan por hacer accesible y navegable un sitio web desaparecido o modificado y se encargan de seleccionar, programar y verificar las recolecciones- son, en muchas ocasiones, autores de entradas en blogs, artículos en revistas especializadas y de comunicaciones en encuentros internacionales que reflexionan sobre los desafíos y los logros del web *archiving* (AlSum, 2016)

De hecho, abundan los escritos de quienes estaban involucrados en los proyectos de conservación del patrimonio documental de la Red, tanto a nivel internacional, como dentro las fronteras nacionales (véase, como ejemplos: Barone, Zeitlyn y Maer-Schönberger, 2015; Goldman, 2015). Gran parte de los avances, pues, se basan en la cooperación entre servicios tecnológicos de las diferentes bibliotecas patrimoniales. Esta cooperación se demuestra periódicamente a través de los encuentros entre técnicos y conservadores web de diversas bibliotecas europeas para discutir sobre el desarrollo del software compartidos que se utilizan para organizar y recolectar los sitios web⁸⁰⁹. En las redes sociales, es posible encontrar una comunidad virtual de archivadores del web como, por ejemplo, los que están detrás del *hashtag* #webarchiving en Twitter.

Si, por lo general, los escritos tienden a presentar la actividad de archivo de la web como algo novedoso, necesario y en continuo desarrollo, otros han destacado el estancamiento de la reflexión y del desarrollo de esta tarea de los bibliotecarios (Miller, 2015). Por ejemplo, Kalev Leetaru (2017), ha puesto de manifiesto que, en los últimos congresos sobre esta materia, las charlas eran idénticas entre ellas, se repetían los temas de los años anteriores y no hubo progresos relevantes. Además, ha denunciado las tendencias de los conservadores web a citar el número de URL recolectadas en vez de hablar del número de páginas o artículos recolectados, que suelen contener un número de URL mucho mayor, como imágenes, contenidos embebidos, etc. De hecho, ha

⁸⁰⁹ Véase el workshop organizado este año en Viena en la página: <https://sbforge.org/display/NAS/2017+NAS+workshop>.

afirmado que el mayor peligro de los archivos web es caer en la trampa de la mera recolección, sin prestar atención a la necesidad de monitorear las novedades en la Red y verificar la calidad de los contenidos ya archivados. Para mejorar estas dos actividades, la participación de los usuarios y, en particular, de los historiadores, podría ser determinante.

Finalmente, una de las ramas más interesantes de los estudios sobre el archivado web es la que se ocupa de cómo los investigadores pueden aprovechar esta nueva herramienta. Existen numerosos blogs dedicados a cómo aprovechar los archivos de la web para nuevas investigaciones. En ellos se reflexiona sobre las cuestiones metodológicas vinculadas a unas fuentes hasta hace pocos años casi totalmente desconocidas. En esta línea, uno de los autores más influyentes en el campo de la historiografía y del uso y de la interpretación de los *big datas* es Ian Milligan (2017). Este y otros historiadores han especulado sobre los avances de la historia digital y afirman que nos encontramos en una fase de transición para la historia del presente (Graham, Milligan y Weingart, 2015). Asimismo, esto implica una reflexión más profunda sobre la diferencia, que destaca Anaclet Pons (2011), entre la «digitalización de las fuentes conservadas bajos otros soportes» y los «recursos cuyo origen es propiamente digital: el patrimonio *born digital*». Si las primeras presentan unas potencialidades y unas problemáticas relativamente conocidas, la conservación y la consulta de las fuentes «nacidas digitales» plantean nuevos retos a los investigadores.

En España, los historiadores familiarizados con el entorno digital conocen los archivos web, aunque su uso y su análisis no han tenido gran difusión. Para quedarnos en las propias mesas dedicadas a la historia digital de los congresos AHC en 2014 y 2016, las referencias a Internet Archive han sido marginales y anecdóticas (García Algarra, 2015: 5877; Eiroa, 2017: 3157). También al último número de la revista *Ayer* con un dossier sobre este tema, sólo un artículo cita, y de pasada, la existencia de esta herramienta (Eiroa, 2018: 94).

Para concluir esta breve revisión sobre los estudios en materia de archivado web e historia digital, increíblemente numerosos, podemos mencionar el proyecto inglés llamado *Big UK Domain Data for the Arts and Humanities* (más conocido como BUDDAH Project) para aprovechar los *big datas*, para las investigaciones en humanidades e historia del arte, compuesto por diversas universidades y bibliotecas inglesas y danesas⁸¹⁰.

Los recursos necesarios para gestionar un archivo web. La OpenWayback

En este capítulo nos centraremos en las herramientas utilizadas en la BNE para la gestión de las colecciones. En particular, analizaremos la herramienta de acceso a la colección, la OpenWayback, que resulta fundamental, tanto para la labor de los conservadores web, como para la visualización de los contenidos archivados abierto a los usuarios. Sin embargo, hay que mencionar otros recursos tecnológicos utilizados para realizar el Archivo de la Web Española (AWE), que son esencialmente dos: NetarchiveSuite (NAS) y Colecciones en la Web (CWeb).

NAS es un software *open source* que ha sido creado para archivar páginas web con una interfaz segura pensada para los bibliotecarios y los conservadores del archivo web. Gracias a NAS, adoptado por la BNE en 2014, es posible programar y recolectar enormes cantidades de datos

⁸¹⁰ Las informaciones sobre este proyecto se encuentran en: <https://buddah.projects.history.ac.uk/>.

contenidos en la web. El software utiliza un robot, un web *crawler* llamado Heritrix⁸¹¹, desarrollado por Internet Archive, que es capaz de acceder como un usuario a decenas de sitios web cada hora y de rastrear su contenido y su estructura (Sørensen, 2017). Además, NAS permite verificar el resultado del rastreo a través de los datos concretos de las colecciones activas: el periodo de rastreo, el número de documentos recolectados y los errores encontrados por Heritrix, entre otras funciones (véase Rosenthal, 2016). Asimismo, existen contenidos que no se consiguen recolectar, como la *dark web*, es decir, aquellos contenidos a los que se accede por búsqueda, las bases de datos, los elementos en *streaming*, y todos los protocolos diferentes de http, que constituyen un reto por la comunidad de técnicos y conservadores de la web.

CWeb es una herramienta de gestión de las colecciones plenamente vinculada a NAS. Creada por la Bibliothèque Nationale de France, la BNE obtuvo gratuitamente su uso en 2014. CWeb tiene una interfaz mucho más amigable que NAS y permite una configuración más rápida y precisa de los recursos que se quieren archivar. Sin embargo, hay que transferir a NAS todos los cambios introducidos en CWeb para que tengan efecto en los lanzamientos sucesivos. Aun así, CWeb es particularmente útil para visualizar los listados de las colecciones y elegir los parámetros de los registros de cada semilla o recurso de las llamadas recolecciones selectivas (DPEL, 2017).

Aparte de las cuestiones vinculadas a las exigencias éticas y legales de conservación e investigación, hay que tener en cuenta las modalidades de las recolecciones. Existen dos tipos principales de recolección: la masiva y la selectiva. Con la primera, se pretende fotografiar la red, aunque de una forma más superficial y con una frecuencia media de un año o un semestre, con el intento de reflejar lo que se ofrece en una determinada época bajo el mismo dominio. La última recolección masiva ha permitido el almacenamiento completo del 87% de los dominios registrados en Red.es (BNE, 2018). Desde luego, el dominio de primer nivel, por ejemplo, .es, no representa el conjunto del patrimonio documental on-line de un país, ya que muchos recursos se encuentran bajo dominios diferentes (.com, .org, .net, .tk, etc.).

Las recolecciones selectivas, en cambio, se centran en un número extremadamente más reducido de documentos, pero pretenden restituir una capacidad de navegación más parecida a la web viva. Las recolecciones selectivas pueden ser «colecciones de eventos», cuando aspiran a mostrar la reacción de la sociedad, es decir, su reflejo en Internet, ante un evento específico, como unas elecciones, un atentado, un desastre medioambiental, un acontecimiento de relevancia nacional o local. Se trata de colecciones especiales que incluyen páginas de blogs, noticias internacionales, redes sociales, etc. (BNE, 2017)⁸¹².

Asimismo, pueden consistir en «colecciones estables», normalmente temáticas, que están dotadas, a menudo, de rastreos periódicos: es el caso de las recolecciones diarias de la prensa en línea, de las recolecciones semanales de los blogs de comentaristas políticos; de los rastreos mensuales de las páginas universitarias; etc. (Rosenthal, 2016). Los sitios web que tienen la que podríamos definir una «esperanza de vida» más corta merecen especial atención. Barone, Zeitlyn y Maer-Schönberger (2015) han clasificado los sitios que pueden tener un riesgo mayor de desaparecer: por un lado, los sitios académicos o científicos, generalmente vinculados a proyectos determinados o a financiaciones externas, y las páginas personales, de políticos o periodistas; por

⁸¹¹ Heritrix complementa el trabajo de Alexa, otro robot de Internet Archive. Véase el artículo en New Tech News disponible en: <http://www.newtechnews.co.uk/heritrix-web-crawler> [Consulta: 29/08/2017].

⁸¹² Una de las últimas recolecciones lanzadas es la que se ha dedicado a los atentados de Barcelona y Cambrils.

otro, los blogs y las plataformas sociales que tienen una «esperanza de vida» diferente de las páginas institucionales, como la de los organismos públicos o de los periódicos.

La BNE se ocupa de las colecciones temáticas (Política Nacional, Bellas Artes, Universidades, etc.); de las colecciones dedicadas a un evento específico (la abdicación del Rey, las elecciones generales, etc.); y de las colecciones de riesgo, es decir, de los sitios web relevantes próximos a la desaparición. Pese a que la iniciativa del Depósito pertenezca a la BNE, cada comunidad autónoma posee un centro de conservación de las colecciones regionales.

La herramienta de acceso a los recursos archivados por la BNE se llama *OpenWayback*. Las *wayback* son los softwares utilizados para visualizar los contenidos archivados por los diferentes centros de conservación en todo el mundo. Se trata, en realidad, de las interfaces necesarias para orientarse en los vastos contenidos recolectados, que evitan que la búsqueda lleve al recurso original, pero permiten realizar una navegación parecida a la que se realizaría en la web viva. La más famosa, la Wayback Machine de Internet Archive, ha sido también la más analizada. De hecho, los conservadores y los usuarios se han interrogado sobre la calidad de los archivos gracias a estas herramientas (Leetaru, 2015). Aunque se pueda navegar por las páginas, éstas aparecen «congeladas»: los foros de discusión no permiten nuevas entradas, no es posible usar los motores de búsquedas ni acceder a las bases de datos o visualizar un contenido en *streaming*.

Para visualizar y navegar por las páginas web archivadas se utiliza un software *open source* llamado OpenWayback (OW). Como otras herramientas de este estilo, OW permite moverse fácilmente hacia una versión anterior o posterior del recurso visualizado. Esta función es esencial para que el usuario pueda tener una rápida visión de los cambios en el tiempo que ha sufrido una página web. La búsqueda por URL, de momento, es la única habilitada para encontrar los recursos. Pero, este tipo de búsqueda limita mucho la posibilidad de encontrar recursos que o no se conocen o no se recuerdan o pueden haber cambiado de URL a lo largo del tiempo. No es posible, de momento, efectuar búsquedas a texto completo, ya que no existe un motor, como google, capaz de encontrar una palabra o una frase en todo lo archivado o en una parte de ellos.

También la visualización de los recursos puede resultar frustrante. La experiencia de navegación en la OpenWayback pretende imitar la que se realizaría en la web viva, tal y como el recurso aparecía el día de su captura. Esto quiere decir que se quieren respetar las modalidades de acceso, navegación y visualización tanto del texto, como de las imágenes, de los vídeos, de los elementos embebidos, de los hiperenlaces, etc. No obstante, el sitio web capturado tiene, por lo general, límites de navegación debidos a una configuración de rastreo que no permite adentrarse en todos los niveles del propio sitio: tras un par de clic es fácil encontrar enlaces que llevan a recursos no archivados.

Además, Heritrix, a menudo, no consigue recolectar algunos recursos, como los audios y los videos, los contenidos embebidos, las bases de datos, etc. Finalmente, la misma OpenWayback está plagada de pequeños o grandes errores de visualización -debidos, sobre todo, a la incompatibilidad con el diseño o la programación de la página, por ejemplo, por el uso de Javascript o Flash- lo que, en ocasiones, impide la correcta navegación. Asimismo, existen algunos límites legales a la captura y a la consulta de los contenidos de la Web. De hecho, el Real Decreto de 2015, ya mencionado, establece también que el Depósito legal de las publicaciones en línea tiene que respetar los límites impuestos por la Ley de Propiedad Intelectual⁸¹³ (González, 2015).

⁸¹³ Disponible en: <https://www.boe.es/buscar/act.php?id=BOE-A-1996-8930>.

Esto quiere decir que las instituciones que efectúan rastreos masivos tienen más inconvenientes a la hora de dejar libre el acceso al archivo web.

El AWE ha abierto el acceso al público en el verano de 2017. En una primera y provisional fase el acceso ha quedado restringido a algunos terminales de la propia BNE y de las bibliotecas habilitadas por la Comunidades Autónomas. Por ello, quedan prohibidas las impresiones, las copias electrónicas y las fotografías. Esto no implica que no sea posible saber lo que se está recolectado en el AWE. La búsqueda en el OpenWayback es disponible desde cualquier terminal con conexión a Internet⁸¹⁴. A través de un largo listado de sitios web recolectado es posible efectuar una búsqueda por URL. Ésta lleva a un calendario que señala los días en los que este recurso se capturó. Sin embargo, como se ha dicho, la visualización de los recursos recolectados sufre de unas severas restricciones de acceso, para respetar la legislación sobre propiedad intelectual: una vez verificada la presencia de los contenidos de interés, hay que dirigirse a los centros con ordenadores habilitados.

Creación y seguimiento de una colección temática

Para proponer, iniciar y mantener una nueva colección no hay que tener en cuenta gastos excesivos para las herramientas hardware y software. No obstante, introducir numerosas semillas puede significar un aumento de las horas de trabajo por persona a la vez que del espacio ocupado en los servidores. En este último capítulo, nos centraremos en un ejemplo práctico para explicar la tarea de selección y mantenimiento de los sitios web. No hablaremos sólo de las razones detrás de la creación de una colección temática y de la selección de los recursos, sino de la configuración concreta de sus semillas. Esto nos permitirá reflexionar sobre las tareas de conservación del web y el control de calidad de los recursos archivado y de cómo esto podría influir en la percepción del servicio por parte de los usuarios.

La selección de los sitios web, aquí, limitada a unos pocos ejemplos, tiene que tener en cuenta la relevancia de los recursos para la sociedad y la cultura española. No existe una fórmula definida y exacta para elegir las llamadas semillas, es decir las URL elegidas para ser rastreadas. La selección, pues, depende de las competencias, del tiempo y del esfuerzo empleado por los conservadores web, así que una colección puede resultar más o menos completa, aunque siempre quedarán agujeros y serán necesarias nuevas actualizaciones.

En cambio, la configuración de cada semilla a pesar de tener una componente de arbitrariedad, intenta respetar unos criterios uniformes según lo que se pretende recolectar. Esta tarea se centra, esencialmente, en tres parámetros principales: el tamaño, la frecuencia y la profundidad. No existe una combinación perfecta aplicable a todos los sitios web, así que hay que entender cuál es el criterio utilizado por los conservadores a la hora de acercarse a un recurso. El término *tamaño* puede llevar a confusión: se puede referir al número de bytes que el robot está autorizado a recolectar o al límite del número de URL. Esta segunda opción es la que se puede seleccionar en CWeb, que permite elegir entre tamaño pequeño (hasta 10.000 URLs), mediano (hasta 50.000 URLs) y grande (hasta 100.000 URLs).

⁸¹⁴ El OpenWayback del Archivo de la Web Española se encuentra en: <http://www.dl-e.es/openwayback/wayback/>.

La *frecuencia* es el parámetro más visible para un usuario. El calendario de OpenWayback muestra de un vistazo el número de veces que la URL buscada ha sido archivada. Los conservadores pueden elegir, en CWeb, entre una frecuencia única o periódica: diaria, semanal, mensual, semestral y anual. La configuración de este elemento es determinante para reflejar las evoluciones de una página web, pero tiene que adecuarse al ritmo de actualización de la misma. No tiene sentido recolectar diariamente un sitio web que tiene un número reducido de nuevas entradas mensuales. Por otro lado, en la BNE, se efectúa un rastreo diario de todas aquellas páginas web que se actualizan frecuentemente y que tienen una relevancia especial. Por ejemplo, un periódico on-line suele necesitar de una configuración de la frecuencia como diaria, para conseguir capturar los continuos cambios en la portada y la introducción de nuevos artículos.

Finalmente, con el término *profundidad* nos referimos al número de clics que, a partir de la URL principal, son necesarios para llegar a un recurso. En CWeb, los conservadores pueden elegir entre varias posibilidades: Inicio, Inicio+1, Inicio+2, Host, Ruta y Dominio. Por ejemplo, de la portada de un periódico (nivel Inicio, útil por si se quiere recolectar un recurso sin enlaces) es posible acceder a un segundo nivel (Inicio+1), es decir a todas las secciones que se encuentran en el menú y a todos los artículos que se encuentran en ella. A partir de este según nivel, se puede acceder a otros recursos que no sería posible alcanzar desde la portada (Inicio+2). Si se quisiera evitar interrupciones en la cadena de hiperenlace dentro de un sitio web, se elegiría la profundidad Ruta (para recolectar todo un directorio) o Host (para un sitio web completo). Con la profundidad Dominio se recolectarían todos los sitios web vinculados a un determinado dominio, lo que sería una exageración para las páginas web reunidas bajo un dominio común, como wordpress.com, pero resulta de gran utilidad para recolectar un sitio pequeño o mediano en riesgo de desaparición. Si los usuarios pueden percatarse de la frecuencia seleccionada para un sitio web a través del calendario de OpenWayback, para verificar la profundidad deben consultar una hoja Excel preparada por los conservadores web de la BNE⁸¹⁵.

El caso concreto que se ha decidido proponer en este texto, y que se presenta a continuación, sirve para reflejar la labor de configuración de las semillas que intenta alcanzar un equilibrio entre los recursos disponibles y el afán de conservación.

Nos centraremos en una colección de los contenidos en línea sobre la historia contemporánea de España y, en particular, de las páginas personales, blogs y cuentas en las redes sociales de los más importantes historiadores contemporaneístas. Podríamos imaginarlos como una extensión digital de sus publicaciones académicas y en la prensa, con un registro más divulgativo. Pese a que «las comunidades académicas ligadas a las ciencias sociales permanecen, hasta hoy, a la vera de la interacción online», algunos historiadores intervienen, a menudo, en las cuestiones historiográficas a través de sus páginas web (Quiroga, 2011: 77). Reflexionan sobre la memoria colectiva y el debate político, aportando unas especulaciones peculiares y poco visibles en otros medios de comunicación, como la televisión o la radio. Pero no se trata sólo de doctores universitarios, con entradas de divulgación científica, sino también de profesores de institutos que proponen pistas para la docencia y clases sobre temas específicos.

También hay sitios web de los organismos públicos (como el portal gubernamental dedicado a la Ley de Memoria Histórica), de algunas asociaciones de historiadores (como la AHC) o de fundaciones (Mapfre); páginas de humor sobre la historia de España; sitios de recolección de

⁸¹⁵ Las hojas Excel de las colecciones selectivas y autonómicas se encuentran en el sitio de la BNE respectivamente en: <http://www.bne.es/es/Colecciones/ArchivoWeb/Subcolecciones/selectivas.html> y <http://www.bne.es/es/Colecciones/ArchivoWeb/Subcolecciones/ComunidadesAutonomas.html>.

podcast dedicados a la historia, etc. Es de destacar que durante la búsqueda se han encontrado numerosos sitios web que ya han desaparecido, como Antena Historia o Reportero de la Historia⁸¹⁶, lo que nos señala la pérdida continua de parte del patrimonio digital español.

El caso concreto que se ha decidido proponer en este texto, y que se presenta a continuación, sirve para reflejar la labor de configuración de las semillas que intenta alcanzar un equilibrio entre los recursos disponibles y el afán de conservación.

En las tablas, que presentan sólo algunos de estos ejemplos sin afán de exhaustividad, hemos señalado algunos datos necesarios para crear el registro en CWeb. Aparte de los parámetros arriba mencionados (frecuencia, profundidad y tamaño), del nombre y del URL principal, hemos añadido, los URL adicionales, en caso de niveles difíciles de alcanzar con el número de clics configurado y que necesitan de una recolección especial. Si se configura, por ejemplo, una profundidad de Inicio+1, el robot rastreará no sólo a partir de la Semilla o URL principal sino también todos los enlaces vinculados a esas URL adicionales. Es una forma de garantizar que se archiven los contenidos que se consideren relevantes. Además, se proponen unas palabras claves -respetando el catálogo de autoridades de la BNE- que puedan identificar el sitio web y facilite, en un futuro, las indizaciones y las búsquedas a texto completo.

Las diversas combinaciones entre frecuencia y tamaño crean diferentes *harvest*: esto quiere decir que los datos sobre las recolecciones se dividen según la configuración de las semillas. Miremos los ejemplos. Las recolecciones mensuales se dividirían en dos cosechas o *harvests* (Tabla 1 y Tabla 2). Eso quiere decir que el lanzamiento sería simultáneo para cada semilla que perteneces a cada uno de ellas.

Tabla 1. Ejemplo de *harvest* con frecuencia Mensual y tamaño Pequeño

Título	Semilla	Profundidad	Palabras clave
Los archivos de Justo Serna	https://justoserna.com/	Inicio+1	Blogs / Historiadores / Serna, Justo 1959-
Aula de Historia Contemporánea - Blog de José Manuel Fernández Ros	http://jomros.blogspot.com.es/	Inicio+1	Blogs / Historia Contemporánea / Enseñanza
Historia en Meme	https://historiaenmeme.tumblr.com/	Inicio+1	Blogs / Historia de España / Sátira

Cada *harvest* puede contener más de un *job* -es decir, dos trabajos de recolección en la misma cosecha- cuando hay semillas con diversa profundidad: Inicio+1 e Inicio+2 (Tabla 2 y Tabla 4). Asimismo, se puede notar que, para el blog de Julián Casanova, hemos propuesto un par de URL

⁸¹⁶ Las URL llevan a unas páginas que señalan que el dominio ha expirado: <http://www.reporterodelahistoria.com/> o más claramente: <https://www.blogger.com/blogin.g?blogspotURL=http://antenahistoria.blogspot.com.es/>.

adicionales, para que no se pierdan aquellos contenidos, como sus artículos, que se encuentran a dos clics de la página inicial.

Tabla 2. Ejemplo de *harvest* con frecuencia Mensual y tamaño Mediano

Título	Semilla	Profundidad	Palabras clave
Julián Casanova. La historia y sus sentidos	http://www.juliancasanova.es/ + URL adicionales: http://www.juliancasanova.es/articulos/ ; http://www.juliancasanova.es/articulos-3/	Inicio+1	Blogs / Historiadores / Casanova, Julián
La factoría histórica	https://factoriahistorica.wordpress.com/	Inicio+1	Blogs / Historia Contemporánea
El sitio web de la historia del siglo XX	http://www.historiasiglo20.org/	Inicio+2	Blogs/Historia Contemporánea/Enseñanza
Memoria Histórica	http://www.memoriahistorica.gob.es/es-es/Paginas/index.aspx	Inicio+2	Gobierno de España / Ley de Memoria Histórica
Asociación de Historia Contemporánea	https://www.ahistcon.org/	Inicio+2	Asociaciones profesionales / Historia Contemporánea / Historiadores

También en el caso de las semillas recolectadas de forma semanal se podrían crear dos *harvests*, según el tamaño configurado. Las cuentas de Twitter, como sugiere el manual para conservadores web, deben tener una frecuencia semanal (DPEL, 2017). Aun así, si se trata de una cuenta especialmente activa en un momento determinado, se podría optar por una frecuencia diaria.

Tabla 3. Ejemplo de *harvest* con frecuencia Semanal y tamaño Pequeño

Título	Semilla	Profundidad	Palabras clave
Twitter de Justo Serna	https://twitter.com/justoserna?lang=es	Inicio+1	Redes sociales / Historiadores / Serna, Justo 1959-
Twitter de Julián Casanova	https://twitter.com/CasanovaHistory?lang=es	Inicio+1	Redes sociales / Historiadores / Casanova, Julián
Mirada Histórica - Blog de José M. Fernández Ros	https://miradahistorica.com/	Inicio+1	Blogs / Historia Contemporánea / Enseñanza
Biombo Histórico - Blog de Vicente Camarasa	https://biombohistorico.blogspot.com.es/	Inicio+1	Blogs / Historia Contemporánea / Enseñanza
Metahistoria	http://metahistoria.com/	Inicio+1	Blogs / Historiadores / Historia de España

Se puede notar que, en el caso de los sitios web con contenidos pesados, como videos o audios, hemos preferido una configuración del tamaño mediano.

Tabla 4. Ejemplo de *harvest* con frecuencia Semanal y tamaño Mediano

Título	Semilla	Profundidad	Palabras clave
Proyecto Clío. History and History Teaching	http://clio.rediris.es/	Inicio+2	Blogs / Historia Contemporánea / Didáctica
Memorias de un Tambor	https://memoriasdeuntambor.com/	Inicio+1	Blogs / Historia de España / Podcast
Histocast	http://www.histocast.com/	Inicio+1	Blogs / Historia de España / Podcast

Finalmente, hemos considerado los casos de los sitios web que aún existen, pero han dejado de actualizarse. Probablemente, se trata de los ejemplos más claros de la aportación del archivado web, ya que estas páginas están en riesgo de desaparecer.

Tabla 5. Ejemplo de *harvest* con frecuencia Única y tamaño Mediano

Título	Semilla	Profundidad	Palabras clave
Proyecto Clío. Haciendo historia en la era digital...	http://proyectoclio.blogspot.com.es/	Ruta	Blogs / Historia Contemporánea / Didáctica
Historia. Fundación Mapfre	http://historia.fundacionmapfre.org/historia/es/	Ruta	Historia de España / Mapfre

Una vez creado los registros de las semillas en CWeb y transferidos los cambios a NAS, el robot rastreará cada *harvest* el día seleccionado, para la recolección única, y periódicamente cada día de la semana o del mes configurado. Desde entonces, los conservadores tienen que hacer el seguimiento de la colección, lo que supone no sólo una continua revisión de las semillas seleccionadas, sino verificar, en NAS, que el robot no haya encontrado errores y, en OpenWayback, que la captura respete la apariencia y la interactividad de la web viva.

También las aportaciones de los propios usuarios pueden resultar determinante. La apertura al público del AWE es demasiado reciente para verificar el nivel de implicación y satisfacción de los investigadores u de otros interesados. Comprendiendo la labor de los conservadores web, los investigadores pueden percibir las potencialidades de esta nueva fuente histórica y tener una idea más clara de lo que se pueden encontrar en un archivo web y del porqué haya tantos fallos.

Por otro lado, a través de una colaboración entre usuarios y conservadores web, estos últimos afinarían su capacidad de seleccionar y configurar las semillas de interés, según los intereses de quienes luego las aprovecharían como recursos para la investigación. Por esta razón, los usuarios tienen que ser considerado unos recursos imprevisibles pero innovadores que, a través de

sugerencias, *feedback* o la señalación de los errores y faltas que detecten, pueden contribuir a mejorar el archivado del patrimonio digital del país.

Conclusiones

Las informaciones, no sólo textuales, contenidas en la Web representan la memoria de las organizaciones y de las comunicaciones de millones de personas que, gracias a esta herramienta, han creado las llamadas «comunidades virtuales». Éstas son agregaciones sociales que a veces no son el reflejo del mundo real, pero pueden salir de las fronteras territoriales, re-configurar las tradicionales formas de sociabilidad y determinar varios aspectos de la vida individual y colectiva (como la movilización política coyuntural o la difusión mundial de contenido artístico-cultural).

Ahondando en esa idea, podríamos decir que los sitios web constituyen el frágil patrimonio digital de la humanidad entera. Esto no quita que las poblaciones de cada Estado se consideren «subcomunidades virtuales» con una identidad y caracteres más o menos homogéneos. De hecho, la pérdida, por sustitución o desaparición, de los contenidos de Internet, con el notorio mensaje «404 not found», va en detrimento del patrimonio histórico y cultural de un país. Por esta razón, y a pesar de las perplejidades alrededor de las dimensiones de la Red, resulta ya irrenunciable la necesidad de unas políticas para generar y mantener un archivo nacional de los sitios web a disposición de los investigadores u otros usuarios presentes y futuros.

Internet Archive no se ha considerado una respuesta satisfactoria, ni por parte de los bibliotecarios, ni por parte de los investigadores. Por esta razón, la acción de los centros de conservación y, sobre todo, la colaboración entre ellos se ha revelado imprescindible para encontrar una solución permanente al problema de la pérdida de contenidos en internet. «We have to find a way to make it an official part of the coding process», ha afirmado un periodista de tecnología (Miller, 2015).

El Archivo de la Web española, el ejemplo más cercano, no es sólo un robot que, automáticamente, recolecta cualquier contenido de la Web, ni una entidad abstracta, centralizada y estandarizada. Hay seleccionadores y curadores detrás de muchas elecciones y, además, cuestiones técnicas y legales que las justifican.

En esta línea, los historiadores, como los antropólogos y los sociólogos y los expertos de otras disciplinas, por sus competencias en el uso de la documentación y por sus inquietudes, podrían contribuir a identificar los recursos más significativos. La colaboración entre usuarios y conservadores web reduciría la dificultad en la selección las páginas que podrían tener una cierta relevancia para futuras investigación.

Desde luego, esto se vincula a la necesidad de mejorar la interfaz de las herramientas de acceso, como OpenWayback, y la indización de los recursos para facilitar las búsquedas a texto completo. No se trata sólo de conservar los contenidos más visibles de la web y la versión histórica de las páginas. La investigación académica, como ha señalado Mar Pérez (2014:145), puede «desarrollar minería de datos o enlaces a gran escala que puede ayudar a extraer patrones y tendencias». Además, el archivado web abre nuevas posibilidades para entender cómo el mundo, una comunidad virtual o unas páginas locales han reaccionado ante los retos contemporáneos. Pero, para que los investigadores se acerquen a esta nueva herramienta, deben conocer cómo esta

funciona, cuáles son los contenidos que se pueden encontrar y por qué a veces los recursos no aparecen tal y como se desearía. Como ha destacado Leetaru (2016), conocer cómo funciona el archivado de la web nos permite entender cuáles son las lógicas que están detrás de la selección y visualización de los contenidos que un día usaremos como fuentes para analizar la sociedad del siglo XXI.

Bibliografía

- ALSUM, Ahmed, 2016. Web Halloween, secrets of resurrecting the SLAC dead website from the computer cemetery to live Web. En: *Digital Library Blog* [en línea]. Disponible en: <http://library.stanford.edu/blogs/digital-library-blog/2014/10/web-halloween-secrets-resurrecting-slac-dead-website-computer> [consulta: 25/08/2017].
- DPEL (Área de Gestión del Depósito de las Publicaciones en Línea), 2017. *Manual para conservadores web*. [documento interno]. Madrid: Biblioteca Nacional de España. Versión 09.
- BANOS, Vangelis, 2015. *Web Crawling, Analysis and Archiving* [tesis doctoral]. Salonicchi: Department of informatics, Aristotle University of Thessaloniki. Disponible en: <http://vbanos.gr/phd-thesis/>.
- BARONE, F., ZEITLYN, D. y MAER-SCHÖNBERGER, V., 2015: Learning from failure: the case of the disappearing web site. *First Monday* [en línea]. Chicago: University of Illinois, vol. 20, no. 5. [consulta: 25/08/2017]. ISSN 1396-0466 y DOI: <http://dx.doi.org/10.5210/fm.v20i5.5852>. Disponible en: <http://firstmonday.org/ojs/index.php/fm/article/view/5852/4456>.
- BNE (Biblioteca Nacional de España), 2017. Recolección web sobre los atentados de Barcelona y Cambrils. En: *Biblioteca Nacional de España. Noticias* [en línea]. Disponible en: <http://www.bne.es/es/AreaPrensa/Noticias2017/0824-recoleccion-web-atentados-cataluna.html> [consulta: 08/09/2017].
- 2018. La BNE completa su tercera recolección masiva del dominio.es. En: *Biblioteca Nacional de España. Noticias* [en línea]. Disponible en: <http://www.bne.es/es/AreaPrensa/noticias2018/0625-tercera-recoleccion-masiva-dominio-es.html> [consulta: 25/06/2018].
- BRÜGGER, Niels y SCHROEDER, Ralph, eds., 2017. *The Web as History. Using Web Archives to Understand the Past and the Present*. Londres: UCL Press. Disponible en: <http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf>.
- CHEBBI, Aïda, 2013. *Archivage du Web organisationnel dans une perspective archivistique* [tesis doctoral]. Montreal: Université de Montréal. Disponible en: <https://papyrus.bib.umontreal.ca/xmlui/handle/1866/9203>.
- EIROA SAN FRANCISCO, Matilde, 2017. El tratamiento de las fuentes hemerográficas *on line*: Una propuesta metodológica. En: GONZÁLEZ MADRID, Damián Alberto, ORTIZ HERAS, Manuel y PÉREZ GARZÓN, Juan Sisinio (coord.). *¿La Historia, lost in translation? XIII Congreso de la Asociación de Historia Contemporánea*, Albacete, 21 a 23 de septiembre de 2016. Cuenca: Ediciones de la Universidad de Castilla-La Mancha, pp. 3149-3160. ISBN 978-84-9044-265-4.
- 2018. El pasado en el presente: el conocimiento historiográfico en las fuentes digitales. *Ayer*. Madrid: Asociación de Historia Contemporánea y Marcial Pons, Ediciones de Historia, S. A., 110 (2), pp. 83-109. ISSN 1134-2277.
- GARCÍA ALGARRA, Francisco Javier, 2015. Estructura secreta de la web: cómo acceder a las fuentes no indexadas. En: FOLGUERA, Pilar *et al.* (coords.). *Pensar con la historia desde el siglo XXI: actas del XII Congreso de la Asociación de Historia Contemporánea*. Madrid: UAM ediciones, pp. 5869-5881. ISBN 978-84-8344-458-0.

- GOETHALS, Andrea, OURY, Clément, PEARSON, David, SIERMAN, Barbara y STEINKE, Tobias, 2015. Facing the Challenge of Web Archives Preservation Collaboratively: The Role and Work of the IIPC Preservation Working Group. *D-Lib Magazine* [en línea]. Olio, vol. 21, no. 5/6 [consulta: septiembre de 2017]. Disponible en: <http://www.dlib.org/dlib/may15/goethals/05goethals.html>.
- GOLDMAN, Ben, 2015. Web Archiving in Libraries and Archives (and at Penn State). En: *100 Digital Discoveries* [en línea]. Disponible en: <http://sites.psu.edu/100digidisc/2015/05/01/web-archiving-in-libraries-and-archives-and-at-penn-state/> [consulta 06/08/2017].
- GONZÁLEZ, Lía, 2015. Qué es el depósito legal de las publicaciones en línea. En: *Bibliopos, Información y recursos sobre Biblioteconomía y Documentación* [en línea]. Disponible en: <http://www.bibliopos.es/que-es-el-deposito-legal-de-las-publicaciones-en-linea/> [consulta: 03/09/2017].
- GRAHAM, Shawn, MILLIGAN, Ian y WEINGART, Scott, 2015. *The Historian's Macroscope: Big Digital History*. Londres: Imperial College Press. ISBN 978-1783266371 [el último borrador del texto está disponible en: <http://www.themacroscope.org/2.0/>].
- LEETARU, Kalev, 2015. How Much of the Internet Does the Wayback Machine really archive? En: *Forbes/Tech* [en línea]. Disponible en: <https://www.forbes.com/sites/kalevleetaru/2015/11/16/how-much-of-the-internet-does-the-wayback-machine-really-archive/#3fd42a794469> [consulta: 10/09/2017].
- 2016. The Internet Archive turns 20: a behind the scenes look at archiving the web. En: *Forbes/Tech* [en línea]. Disponible en: <https://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#7cdbf8482e07> [consulta: 11/09/2017].
- 2017. Why are libraries failing at web archiving and are we losing our digital history? En: *Forbes/Tech* [en línea]. Disponible en: <https://www.forbes.com/sites/kalevleetaru/2017/03/27/why-are-libraries-failing-at-web-archiving-and-are-we-losing-our-digital-history/#26c76afc6ecd> [consulta 10/09/2017].
- LYMAN, Peter, 2002. Archiving the World Wide Web. En: *Building a National Strategy for Preservation: Issues in Digital Media Archiving* [en línea]. Washington, DC: Council on Library and Information Resources [consulta: 19/08/2017]. ISBN 1-887334-91-2. Disponible en: <https://www.clir.org/pubs/reports/pub106>.
- MILLER, Ron, 2015. The Internet is Failing the Website Preservation Test. En: *Techcrunch* [en línea]. Disponible en: <https://techcrunch.com/2015/08/27/the-internet-is-failing-the-website-preservation-test/> [consulta: 01/08/2017].
- MILLIGAN, Ian, 2017. Archives Unleashed: A Series of Datathons for Cultural Heritage. En: *Web Archives for Historians. Historian who use, think about, and work with Web Archives* [en línea]. Disponible en: <https://webarchivehistorians.org/2017/05/23/archives-unleashed-a-series-of-datathons-for-cultural-heritage/> [consulta: 29/08/2017].
- MUGUERZA LÓPEZ, Icíar y PÉREZ MORILLO, Mar., 2012. El Archivo WEB de la Biblioteca Nacional de España. *Boletín de ANABAD*. Madrid: Federación española de asociaciones de archiveros, bibliotecarios, arqueólogos, museólogos y documentalistas, vol. LXII, no. 4, pp. 101-112. ISSN: 0210-4164.
- PÉREZ MORILLO, Mar, 2014. Los archivos web: un paso más en la preservación del patrimonio documental. El informe técnico ISO 14873 sobre estadísticas e indicadores de calidad en los archivos web. En: MANO GONZÁLEZ, M., es. *Nuevos instrumentos para la evaluación de bibliotecas: la normativa internacional ISO*. Madrid: Aenor, pp. 125-160. ISBN 978-84-8143-845-1.
- PONS, Anaclet, 2011. «Guardar como». La historia y las fuentes digitales. *Historia crítica*, 43, pp. 38-61. ISSN 0121-1617.
- QUIROGA, Nicolás, 2011. Blogs de historia. Usos y posibilidades. *Historia crítica*, 43, pp. 62-80. ISSN 0121-1617.
- ROSENTHAL, Colin, 2016. NetarchiveSuite Overview. En: *NetarchiveSuite-SBForge Confluence* [en línea]. Disponible en: <https://sbforge.org/display/NASDOC52/NetarchiveSuite+Overview> [consulta: 23/08/2017].
- SØRENSEN, Mikis Seth, 2017. NetarchiveSuite. En: *NetarchiveSuite-SBForge Confluence* [en línea]. Disponible en: <https://sbforge.org/display/NAS/NetarchiveSuite> [consulta: 23/08/2017].

- VITALI, Stefano, 2004. *Passato digitale. Le fonti dello storico nell'era del computer*. Milano: Bruno Mondadori. ISBN: 9788842490401.
- WALL, Matthew, 2015. Does the digital era herald the end of history? En: *BBC News* [en línea]. Disponible en: <http://www.bbc.com/news/technology-32315449> [consulta: 16/08/2017].
- WEBSTER, Peter, 2016. What do we need to know about the archived web? En: *Webstory: Peter Webster's blog* [en línea]. Disponible en: <https://peterwebster.me/2016/04/18/what-do-we-need-to-know-about-the-archived-web/> [consulta: 02/08/2017].
- WEINER, Eric, 2016. Will future historians consider these days the Digital Dark Ages? [podcast]. En: *NPR* [en línea]. Disponible en: <http://www.npr.org/2016/01/04/461878724/will-future-historians-consider-these-times-the-digital-dark-ages?ft=nprml&f=2,3,46> [consulta 10/09/2014].